基于频繁主题集偏好的学术论文推荐算法

李 冉，林 泓

(武汉理工大学 计算机科学与技术学院，武汉 430063)

摘 要：针对学术论文推荐中项目冷启动问题，提出了一种基于频繁主题集偏好的协同主题回归模型。该算法考虑到用户在选择学术论文时对研究热点的偏好，使用频繁主题集代表研究热点，将用户对研究热点的偏好表示成用户对频繁主题集的偏好。首先，通过潜在狄利克雷分布主题模型挖掘得到论文—主题概率分布矩阵，并筛选出论文中概率较高的主题；然后，挖掘出频繁出现的主题集合，并得到论文-频繁主题集矩阵；最后，在预测未知评分时融入用户对频繁主题集的偏好。在 CiteULike 数据集上的实验表明，相比于矩阵分解模型和协同主题回归模型，该算法在召回率、准确率和 RMSE 三个指标上都有所提升。

关键词：论文推荐；主题模型；频繁主题集；

中图分类号：TP301.6 doi: 10.3969/j.issn.1001-3695.2018.02.0141

Academic paper recommendation algorithm based on frequent topic sets preference

Li Ran, Lin Hong

(*College of computer Science & Technology*, *Wuhan University of Technology*, *Wuhan* 430063, *China*)

Abstract: This paper proposed a collaborative topic regression model based on the preference for frequent topic sets to address the item-cold-start problem in academic paper recommendation. The algorithm takes into account the user's preference for research hotspots when selecting academic papers, and uses frequent topic sets to represent research hotspots. So, user's preference for research hotspots is expressed as the user's preference for frequent topic sets. Firstly, the papers-topic probability distribution matrix is obtained through LDA algorithm and filter out the topics with higher probability in the paper. Then, the algorithm mines the frequently-occurring topic sets and gets the relationships between papers and frequent topic sets. Finally, the user's preference for frequent topic sets is used for the prediction of unknown scores. Experiments on CiteULike datasets show that the algorithm improves the recall, accuracy and RMSE over the matrix factorization model and the collaborative topic regression model.

Key words: paper recommendation; topic model; frequent topic sets

0 引言

学术论文推荐是推荐系统的一个应用方向，结合被推荐物品（学术论文）的特点，基于内容、协同过滤等在电子商务领域广泛使用的推荐算法在学术论文推荐中也取得了一定效果。

在基于内容的论文推荐的技术中，常用的是利用 TF-IDF 方法将文档表示成以关键词为维度的特征向量[1]，并由特征向量计算得到文档间的相似度，然后基于用户的历史阅读记录进行论文的推荐[2,3]。但 TF-IDF 方法只能统计文档中单词的词频信息，无法捕捉文档内部以及文档间的统计特征，也不能确定文档的语义特征，从而只能向用户推荐表面内容相似的文章。随着主题模型的提出及其在文本挖掘中发挥的重要作用，逐渐有人将主题模型应用于推荐系统[4]。主题模型可以捕捉到文档内的语义信息，从而发现文档潜在的主题特征，通过文档间主题的相似度给用户基于内容的推荐[5]。此外，大量研究工作也被集中在如何将概率矩阵分解模型(probabilistic matrix factorization，PMF)[6]更好的应用到论文推荐中。Wang 等[7]将 PMF 和基于内容的推荐相结合，提出协同主题回归模型(collaborative topic regression，CTR)，通过潜在狄利克雷分布主题模型(latent Dirichlet allocation，LDA)[8]对 PMF 的项目潜在因子特征向量进行增强。协同深度学习的分层贝叶斯模型[9]对内容信息进行深度表示学习，并对反馈矩阵进行协同过滤，显著提高了已有的技术水平。Lu 等人[10]提出作者—会议—时间—主题模型构建用户的主题特征，结合 LDA 构建的论文的主题特征，分别增强 PMF 中的用户潜在因子特征向量和项目潜在因子特征向量。除此，还有一些推荐算法也将重点集中在挖掘更多种类的信息来

收稿日期：2018-02-11；**修回日期：**2018-04-08

作者简介：李冉（1994-），女，山东鄄城人，硕士研究生，主要研究方向为数据挖掘、推荐算法、深度学习（962019857@qq.com）；林泓（1965-），女，副教授，硕士，主要研究方向为数据挖掘、深度学习、语言处理.

丰富用户和论文的特征[11,12]。顺着这一研究思路，本文也对如何更好的构建论文特征向量进行了探究。

用户在某个研究方向下做研究时，首先需要阅读相关领域下的核心技术论文，以便了解该方向的主要研究内容和关键技术；其次，阅读新发表的论文对用户也是至关重要的，可以帮助用户紧跟学科的发展，并开阔眼界；同时，用户对包含热点主题的论文的关注度往往更高。核心论文往往意味着被该方向下的很多人阅读过，因此在推荐核心论文时，采用概率矩阵分解模型向用户推荐同领域下其他用户阅读的论文，可以使其他用户的观点发挥重要作用，推荐效果良好。但对于发表不久，还没有被阅读过的论文，概率矩阵分解模型则不能发挥作用，即存在项目冷启动的问题，因此需要对论文内容进行分析，以便将其推荐给需要的用户。上述提到的 CTR 及系列论文[13,14]将基于内容的推荐和概率矩阵分解模型相结合，一定程度上缓解了概率矩阵分解模型的冷启动问题。但是 CTR 在发掘研究热点方面的能力不够，尤其是对于新发表的论文，基本上依赖于基于内容的推荐，而不能体现论文中研究热点的价值。

鉴于上述问题，本文提出了基于频繁主题集偏好的协同主题回归模型，在预测未知评分时，对包含频繁主题集的论文给予一定程度的偏重，频繁出现的主题集合通常代表学术研究的热点，从而凸显包含研究热点的学术论文的价值。该模型首先对语料库进行建模处理，得到论文在主题上的概率分布；从而挖掘出频繁出现的主题集合；最后在协同主题回归模型中融入频繁主题集对推荐结果的影响。

1　相关工作

1.1　论文主题挖掘

本文使用 LDA 主题模型对实验数据集进行处理，生成论文-主题概率分布矩阵。LDA 是一个语料库的生成模型，它的基本思想是文档被表现为隐含主题的随机混合[8]。对于语料库中的每篇文档，LDA 定义了如下生成过程：

a)从 Dirichlet 分布 α 中取样生成文档 i 的主题分布 θ_i；

b)从主题的多项式分布 θ_i 中取样生成文档 i 第 j 个词的主题 $z_{i,j}$；

c)从 Dirichlet 分布 β 中取样生成主题 $z_{i,j}$ 对应的词语分布 $\phi_{z_{i,j}}$；

d)从词语的多项式分布 $\phi_{z_{i,j}}$ 中采样最终生成词语 $\omega_{i,j}$；

e)重复上述过程，就产生了文档 i。

该模型的输入是语料库的词袋模型，输出是两个多项式分布的参数，一个是“文档—主题”分布 $\boldsymbol{\theta}$，一个是“主题—词”分布 ϕ。通过学习这两个参数，可以获得每篇论文所涵盖的主题比例等信息。本文采用 Gibbs 抽样法对上述参数进行推断。

1.2　频繁项集挖掘

频繁项集是指那些经常一起出现的物品集合，其中，“频繁”是由设定的阈值（即最小支持度）来衡量的，一个项集的支持度被定义为数据集中包含该项集的记录所占的比例。

Apriori 算法[15]是一种挖掘关联规则的频繁项集的经典算法，使用逐层搜索的迭代方法来产生频繁项集，即通过 $(k-1)$ 项频繁集得到 k 项频繁集，共包含两个步骤。首先，自连接获取候选集，第一轮的候选集就是数据集中的项，而其他轮次的候选集则是由前一轮次频繁集自连接得到。然后，对候选集进行剪枝，将候选集中支持度小于最小支持度的项和其子集包含非频繁集的项剪掉。最终得到频繁 n 项集。

本文使用 Apriori 算法对 LDA 模型产生的文档-主题分布进行频繁项集挖掘，得到经常共同出现的主题集合，和各频繁主题集合在每篇论文中的分布情况。

2　基于频繁主题集偏好的推荐模型

2.1　频繁主题集偏好

LDA 主题模型的基本思想表明，每篇论文都有一个或多个主题，即每篇论文都对应一个主题集。因此由 LDA 模型得到的论文-主题概率分布矩阵，如图 1(a)，通过筛选论文中概率值较高的主题，可将矩阵表示成图 1(b)所示的形式，每一行中值为 1 所对应的主题的集合即该论文所包含的主题集。显然，同一方向下的论文包含相同或相近的主题集，并且对于一个特定的主题集在不同的论文中出现的次数越多，代表在该研究方向下的关注度越高。

	Topic 1	…	Topic t	…	Topic k
Paper 1	$P_{1,1}$	…	$P_{1,t}$	…	$P_{1,k}$
⋮	⋮		⋮		⋮
Paper i	$P_{i,1}$	…	$P_{i,t}$	…	$P_{i,k}$
⋮	⋮		⋮		⋮
Paper m	$P_{m,1}$	…	$P_{m,t}$	…	$P_{m,k}$

（a）论文-主题概率分布

Topic 1	…	Topic t	…	Topic k
0/1	…	0/1	…	0/1
⋮		⋮		⋮
0/1	…	0/1	…	0/1
⋮		⋮		⋮
0/1	…	0/1	…	0/1

（b）论文-主题矩阵

图 1　数据表示

基于上述分析，频繁主题集，即经常出现在同一篇学术论文中的主题集合，在一定程度上反映了某个研究领域下的研究热点。包含研究热点的论文对用户的价值往往更大，区分论文之间在热度上的差别，可以为用户推荐更有价值的论文，因此在构建论文特征向量时，应考虑到频繁主题集的影响。尤其对于阅读量较少的论文，CTR 等算法仅依赖于论文的潜在主题，将其推荐给阅读过相似论文的用户[7]，在此基础上考虑用户对频繁主题集的偏好，能进一步的提高推荐效果。

2.2　算法模型表示

概率矩阵分解模型是推荐算法中一种经典的推荐模型，在学术论文推荐中也有广泛应用，它将用户历史评分矩阵分解成主题空间上的用户、论文特征矩阵。该算法具有高扩展性，可以通过在算法中融合相似性、社交网络[16]等信息约束用户、论文特征矩阵，提高推荐效果。CTR 便是基于该模型融入了论文的内容信息，将用户 i 对论文 j 的预测评分 $\hat{R}_{ij}$ 作了如下定义。

其中， $\boldsymbol{u}_i$ 和 $\boldsymbol{v}_j$ 分别代表用户 i 和论文 j 的特征向量。 $\boldsymbol{\theta}_j$ 是通过 LDA 主题模型挖掘得到的论文 j 在主题上的概率分布向量。

$$\hat{R}_{ij}=\boldsymbol{u}_i^T\boldsymbol{v}_j \tag{1}$$

$$\boldsymbol{v}_j=\boldsymbol{\theta}_j+\boldsymbol{\varepsilon}_j \tag{2}$$

用户特征向量仍旧定义为服从均值为 0 的高斯分布，如式(3)所示，论文特征向量的定义如式(2)所示， $\boldsymbol{\varepsilon}_j$ 为服从均值为 0 的高斯分布，用于平衡用户评分记录和论文内容对论文特征向量的影响。

$$\boldsymbol{u}_i\sim N\left(0\middle|\sigma_u^{\ 2}\boldsymbol{I}\right) \tag{3}$$

$$\boldsymbol{\varepsilon}_j\sim N\left(0\middle|\sigma_v^{\ 2}\boldsymbol{I}\right) \tag{4}$$

根据本文上节分析，频繁主题集在用户对论文的选择上有一定的影响。因此，本文提出了基于频繁主题集偏好的协同主题回归模型，在 CTR 中融入频繁主题集的全局影响因子 $\boldsymbol{P}$ ，提高推荐效果。模型示意图如图 2 所示。

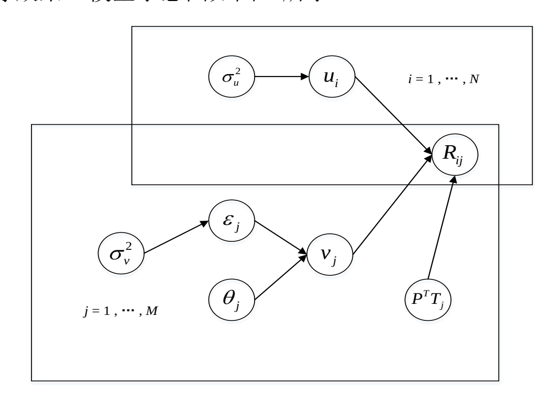


图 2　模型示意图

向量 $\boldsymbol{T}_j=(0/1,0/1,\ldots,0/1)$ 代表论文 j 包含频繁主题集的情况， $\boldsymbol{T}_j$ 的第 s 个值取值为 1，表示论文 j 中含有第 s 个频繁主题集。向量 $\boldsymbol{T}_j$ 的产生过程如下：

a)通过 LDA 主题模型得到论文-主题概率分布矩阵 $\boldsymbol{\theta}$ ；

b)筛选论文中概率值较高的主题，得到每篇论文包含的主题集；

c)使用 Apriori 算法挖掘出频繁出现的主题集合，同时产生矩阵 $\boldsymbol{T}$ 。

为在 CTR 模型中融入频繁主题集的全局影响因子，本文将用户对论文的预测评分重新定义为

$$\hat{R}_{ij}=\begin{cases} g(\boldsymbol{u}_i^T\boldsymbol{v}_j+\dfrac{\boldsymbol{P}^T\boldsymbol{T}_j}{t_j}) & ,t_j\neq 0 \\ g(\boldsymbol{u}_i^T\boldsymbol{v}_j+\dfrac{\boldsymbol{P}^T\boldsymbol{I}}{|\boldsymbol{I}|}) & ,t_j=0 \end{cases} \tag{5}$$

$\hat{R}_{ij}$ 表示预测评分， $\boldsymbol{u}_i$ 和 $\boldsymbol{v}_j$ 的定义同式(3)(2)； $g(x)=1/(1+exp(-x))$ 为逻辑函数，将预测评分映射到[0,1]区间； $\boldsymbol{P}=\left(P_1,P_2,\ldots,P_s,\ldots,P_p\right)$ 是频繁主题集的影响因子向量， P_s 表示频繁主题集 s 在用户对论文评分时产生的影响值， p 是频繁主题集的维度； t_j 表示论文 j 中包含频繁主题集的个数，即向量 $\boldsymbol{T}_j$ 中 1 的个数。当论文 j 中不含任何频繁主题集时，将频繁主题集的影响值定义为所有频繁主题集的影响值的平均值，向量 $\boldsymbol{I}$ 表示单位向量。并且，假定向量 $\boldsymbol{P}$ 和向量 $\boldsymbol{u}$ 和 $\boldsymbol{v}$ 一样服从均值为 0 的高斯分布：

$$p\left(\boldsymbol{P}\,|\,\sigma_p\right)=N\left(\boldsymbol{P}\,|\,0,\sigma_p^2\boldsymbol{I}\right) \tag{6}$$

则可推导出损失函数的定义，如式(7)所示。

R_{ij} 是用户 i 对论文 j 的真实评分； I_{ij} 为指示函数，如果用户 i 对论文 j 有过操作，则返回 1，否则返回 0； λ_u 、 λ_v 和 λ_p 分别为 $\boldsymbol{u}_i$ 、 $\boldsymbol{v}_j$ 和 $\boldsymbol{P}$ 的正则化参数。

通过对向量 $\boldsymbol{u}_i$ 、 $\boldsymbol{v}_j$ 和 $\boldsymbol{P}$ 实施随机梯度下降法，如式(8)所

示，可以求解使损失函数取最小值的用户、论文潜在主题向量以及频繁主题集的影响因子向量 $\boldsymbol{P}$ 的值，从而通过式（1）预测未知评分。

$$E=\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{M}I_{ij}\left(R_{ij}-\hat{R}_{ij}\right)^2+\frac{1}{2}\lambda_u\sum_{i=1}^{N}\boldsymbol{u}_i^T\boldsymbol{u}_i+\frac{1}{2}\lambda_v\sum_{j=1}^{M}\left(\boldsymbol{v}_j-\boldsymbol{\theta}_j\right)^T\left(\boldsymbol{v}_j-\boldsymbol{\theta}_j\right)+\frac{1}{2}\lambda_p\boldsymbol{p}^T\boldsymbol{p} \tag{7}$$

$$\begin{aligned}\frac{\partial E}{\partial u_i}&=\sum_{j=1}^{M}I_{ij}\left(R_{ij}-\hat{R}_{ij}\right)\left(-g'\left(\boldsymbol{u}_i^T\boldsymbol{v}_j+\frac{\boldsymbol{P}^T\boldsymbol{T}_j}{t_j}\right)\right)\boldsymbol{v}_j+\lambda_u\boldsymbol{u}_i,\\ \frac{\partial E}{\partial \boldsymbol{v}_j}&=\sum_{i=1}^{N}I_{ij}\left(R_{ij}-\hat{R}_{ij}\right)\left(-g'\left(\boldsymbol{u}_i^T\boldsymbol{v}_j+\frac{\boldsymbol{P}^T\boldsymbol{T}_j}{t_j}\right)\right)\boldsymbol{u}_i+\lambda_v\left(\boldsymbol{v}_j-\boldsymbol{\theta}_j\right),\\ \frac{\partial E}{\partial \boldsymbol{P}}&=\sum_{i=1}^{N}\sum_{j=1}^{M}I_{ij}\left(R_{ij}-\hat{R}_{ij}\right)\left(-g'\left(\boldsymbol{u}_i^T\boldsymbol{v}_j+\frac{\boldsymbol{P}^T\boldsymbol{T}_j}{t_j}\right)\right)\left(\frac{\boldsymbol{T}_j}{t_j}\right)+\lambda_p\boldsymbol{P}\end{aligned} \tag{8}$$

3　实验结果与分析

3.1　实验方案

本文采用由 CiteULike 网站(http://www.citeulike.org)提供的数据集(http://www.citeulike.org/faq/data.adp)。该数据集包括从 2004 年到 2010 年的 16980 篇论文和 5551 个用户，每个用户都有自己的论文库，其中记录着用户浏览过的论文，共包含了 204986 对用户-论文浏览记录。实验过程中，基于 16980 篇论文依次采用 LDA 主题模型算法和 Apriori 算法，挖掘出频繁出现的主题集合。并且，将每篇论文表示为以频繁主题集合为维度的向量。依次得到矩阵 $\boldsymbol{\theta}$ 和矩阵 $\boldsymbol{T}$，作为预测未知评分时的已知参数。

按照 80%和 20%的比例将用户-论文浏览记录划分为训练集和测试集，进行如下实验：

a)分析频繁主题集的数量、参数 λ_p 对基于频繁主题集偏好的协同主题回归模型的影响，以确定合理的参数值；

b)对比本文模型和 PMF、CTR 的推荐效果。

3.2　评测标准

在评分预测系统中常采用均方根误差（root mean squared error，RMSE）作为度量标准，RMSE 越小，则推荐准确度就越高。RMSE 的求解公式如下，其中 $Test$ 是测试集合。

$$RMSE=\sqrt{\frac{\sum_{(i,j)\in Test}\left(R_{ij}-R_{ij}'\right)^2}{|Test|}} \tag{9}$$

除此之外，推荐系统的目的是向用户推荐用户可能感兴趣的论文，因此，本文在预测用户对论文的评分之后，对用户预测评分进行排序，选取评分分值大且没有被用户操作过的论文推荐给用户，并采用召回率和准确率来衡量推荐效果。假设向用户推荐预测评分最高的 m 篇论文，对于特定用户，其推荐的召回率和准确率定义为

$$recall@m=\frac{TP}{TP+FN} \tag{10}$$

$$precision@m=\frac{TP}{TP+FP} \tag{11}$$

TP 是推荐列表中用户喜欢的论文数量，FN 是没有推荐给用户但用户喜欢的论文的数量，FP 是推荐列表中用户不喜欢的论文的数量。推荐算法的召回率定义为所有用户的推荐召回率的平均值，推荐算法的准确率定义为所有用户的推荐准确率的平均值。

此外，召回率和准确率会出现矛盾的情况，所以经常采用 F-measure 方法去综合考虑两者。F-measure 是召回率和准确率的加权调和平均，特别地，当 $\alpha=1$ 时，就是最常见的 $F1$。本文采用 $F1$ 来衡量推荐效果。

$$F-Measure=\frac{(\alpha^2+1)\cdot precision\cdot recall}{\alpha^2(precision+recall)} \tag{12}$$

3.3　实验结果

3.3.1 参数 p 对推荐效果的影响

在挖掘频繁出现的主题集阶段，当最小支持度设置为不同的值时，得到的频繁主题集的数量也有所不同，反映了当前论文集中的研究热点的分布。设定 LDA 模型的主题个数为 200，最小支持度分别取 0.0014、0.00125、0.00118、0.0012、0.00105，可找出满足这些最小支持度频繁出现的主题集合的数量分别是 54，81，97，118，159。表 1 给出了在推荐列表长度 k 不同的情况下，模型的平均召回率随频繁主题集数量的变化而呈现的不同值。RMSE 的变化趋势，如图(3)所示。实验中的其他参数的设置分别为 $\lambda_u=0.1$、$\lambda_v=0.1$、$\lambda_p=1$。

表 1　频繁主题集数量不同时召回率的对比

推荐列表长度	频繁主题集数量				
	P=54	P=81	P=97	P=118	P=159
k=200	0.7743	0.7901	0.7851	0.7706	0.7538
k=150	0.6909	0.7087	0.6912	0.6895	0.6701
k=100	0.5855	0.5978	0.5782	0.5768	0.5649

k=50	0.4241	0.4336	0.4188	0.4178	0.4068
k=10	0.1663	0.1757	0.1658	0.1654	0.166

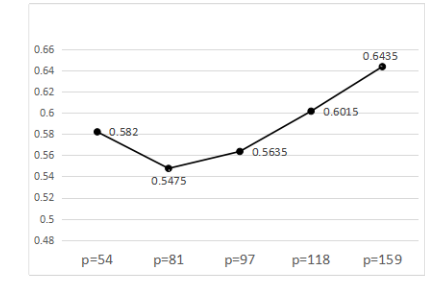


图 3　p 对 RMSE 的影响

频繁主题集的数量越多，RMSE 先随之减小，召回率也相应上升，算法性能提高；但频繁主题集的数量超过一定程度，推荐效果有所降低。实验结果表明，在挖掘频繁主题集时，设置合理的最小支持度，获得与研究热点相对应的频繁主题集，可以使本文算法取得最优的推荐效果。

3.3.2 正则化参数 λ_p 对推荐效果的影响

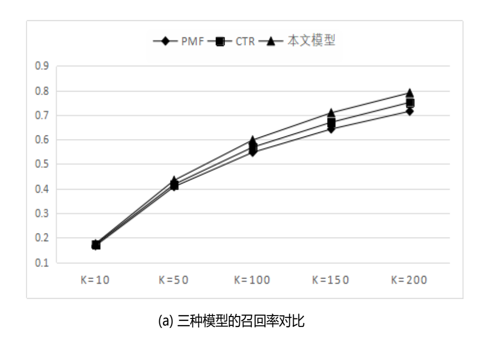
式(7)中的参数 λ_p 越小，在预测未知评分时，用户对频繁主题集的偏好所占的比重越大。为探究频繁主题集对评分的影响，正则化参数 λ_u、λ_v 的设置同上节，选取不同的 λ_p 来衡量 λ_p 对算法性能的影响。实验结果表明，当 $\lambda_p = 1$ 时，本文算法的召回率达到最优，RMSE 的值也较小。

3.3.3 推荐算法比较

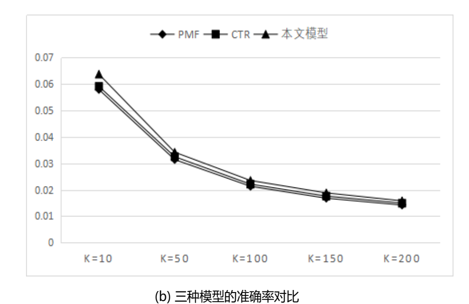
本文模型由原始的 PMF 模型扩展而来，并借鉴 CTR 的思想，与 PMF 和 CTR 模型对比，能够直接体现出本文模型在召回率、准确率和 RMSE 等基准上的提高。因此在本文的实验中，选取了这两种模型作为实验的比较对象。

通过实验，分别得到了使三种模型达到最优效果的参数设置，三种模型的特征空间维度均为 200，PMF 和 CTR 中 $\lambda_u = \lambda_v = 0.01$，本文模型中 $\lambda_u = \lambda_v = 0.1$、$\lambda_p = 1$。在此基础上，设定推荐列表长度 k 分别取{200，150，100，50，10}，对比三种模型在召回率、准确率和 RMSE 上的效果。表 2 展示了详细的实验结果数据，图 4 展现了三种模型在推荐效果上的对比。

实验结果表明，在推荐列表长度不同时，本文模型的召回率、准确率和 F1 都明显优于 PMF 和 CTR，RMSE 的值也有所降低。并且，论文—频繁主题集矩阵 T 可以离线计算，因此本文模型以较小的时间开销代价，获取了推荐效果的提升。



(a) 三种模型的召回率对比



(b) 三种模型的准确率对比

(c) 三种模型的综合测度对比

图 4　三种算法的性能对比

表 2　三种算法的性能对比

指标	算法	k=200	k=150	k=100	k=50	k=10
召回率	PMF	0.7151	0.6426	0.5475	0.4078	0.1673
	CTR	0.7511	0.6701	0.5685	0.4171	0.1725
	本文模型	0.7901	0.7087	0.5978	0.4336	0.1757
准确率	PMF	0.0142	0.0168	0.0213	0.0314	0.0579
	CTR	0.0149	0.0176	0.0221	0.0325	0.0591
	本文模型	0.0158	0.0188	0.0235	0.0342	0.0637
RMSE	PMF			0.6665		
	CTR			0.622		
	本文模型			0.5475		

4　结束语

本文考虑到频繁主题集在用户选择论文时的影响，提出基于频繁主题集偏好的协同主题回归模型，力求帮助用户找到更有价值的学术论文。在真实数据集上的实验证明，基于频繁主题集偏好的协同主题回归模型，对比 PMF 和 CTR 模型，在召回率和准确率上都有一定的提高。

由于用户个性化的需求，频繁主题集的影响值针对不同用户可能不同，因此构建用户敏感的频繁主题集影响向量是下一步的研究重点。

参考文献：

[1] Ramos J. Using TF-IDF to determine word relevance in document queries [C]// Proc of the 1st Instructional Conference on Machine Learning. 2003: 1-4.

[2] Lops P, Gemmis MD, Semeraro G. Content-based recommender systems:

state of the art and trends [M]. New York: Springer, 2011: 73-105.

[3] Philip S, Shola PB, Ovye A. Application of content-based approach in research paper recommendation system for a digital library [J]. International Journal of Advanced Computer Science & Applications, 2014, 5 (10): 37-40.

[4] Krestel R, Fankhauser P, Nejdl W. Latent Dirichlet allocation for tag recommendation [C]// Proc of the 3th ACM Conference on Recommender Systems. New York: ACM Press, 2009: 61-68.

[5] 黄泽明. 基于主题模型的学术论文推荐系统研究 [D]. 大连: 大连海事大学, 2013. (Huang Zeming. Research on recommended system of scholar paper based on topic model [D] Dalian: Dalian Maritime University, 2013.)

[6] Salakhutdinov R, Mnih A. Probabilistic matrix factorization [C]// Proc of the 20th International Conference on Neural Information Processing Systems. USA: Curran Associates Inc. 2007: 1257-1264.

[7] Wang Chong, Blei DM. Collaborative topic modeling for recommending scientific articles [C]// Proc of ACM SIGKDD: International Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 2011: 448-456.

[8] Blei D M, Ng AY, Jordan M I. Latent Dirichlet allocation [J]. J Machine Learning Research Archive, 2003, 3: 993-1022.

[9] Wang Hao, Wang Naiyan, Yeung DY. Collaborative deep learning for recommender systems [C]// Proc of International Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 2015: 1235–1244.

[10] 卢美莲, 张正林, 刘智超. MFWT: 一种推荐学术论文的混合模型 [J]. 北京邮电大学学报, 2016, 39 (4): 24-29. (Lu Meilian, Zhang Zhenglin, Liu Zhichao. MFWT: a hybrid model for academic paper recommender [J]. Journal of Beijing University of Posts and Telecommunications, 2016, 39 (4): 24-29.)

[11] Li Jingming, Cheng Jiaxing, Zhang Wei, *et al.* The research on construction of library website personalized recommendation model based on improved collaborative filtering algorithm [J]. Journal of Changchun Normal University, 2016. 35 (2): 43-48.

[12] Zhang Libin. Research on the application of collaborative filtering technology in the personalized recommendation service for academic resources of university libraries [J]. Hebei Library Journal of Science & Technology, 2017. 30 (4): 85-88.

[13] Wang Hao, Li Wujun. Relational collaborative topic regression for recommender Systems [J]. IEEE Trans on Knowledge & Data Engineering, 2015, 27 (5): 1343-1355.

[14] Wang Hao, Chen Binyi, Li Wujun. Collaborative topic regression with social regularization for tag recommendation [C]// Proc of International Joint Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2013: 2719-2725.

[15] Agrawal R, Imielinski T, Swami AN. Mining association rules between sets of items in large databases, SIGMOD Conference [C]// Proc of ACM SIGMOD International Conference on Management of Data. New York: ACM Press. 1993: 207-216.

[16] Wu Liaoyuan, Jiang Jun, Wang Gang. Study of scientific paper recommendation method based on unified probabilistic matrix factorization in scientific social networks [J]. Computer Science, 2016, 43 (9): 219-223.